Journal of University Teaching and Learning Practice

Engaging Academics with a Simplified Analysis of their Multiple-Choice Question (MCQ) Assessment Results

Geoffrey T. Crisp

University of Adelaide

geoffrey.crisp@adelaide.edu.au

Edward J. Palmer

University of Adelaide

edward.plamer@adelaide.edu.au

Abstract

The appropriate analysis of students' responses to an assessment is an essential step in improving the quality of the assessment itself as well as staff teaching and student learning. Many academics are unfamiliar with the formal processes used to analyze assessment results; the standard statistical methods associated with analyzing the validity and reliability of an assessment are perceived as being too difficult for academics with a limited understanding of statistics. This inability of academics to apply conventional statistical tools with authority often makes it difficult for them to make informed judgements about improving the quality of the questions used in assessments. We analyzed students' answers to a number of selected response assessments and examined different formats for presenting the resulting data to academics from a range of disciplines. We propose the need for a set of simple but effective visual formats that will allow academics to identify questions that should be reviewed before being used again and present the results of a staff survey which evaluated the response of academics to these presentation formats. The survey examined ways in which academics might use the data to assist their teaching and students' learning. We propose that by engaging academics with a formal reflection of students' responses, academic developers are in a position to influence academics' use of specific items for diagnostic and formative assessments.

Introduction

Summative assessments are a high stakes activity for both students and teachers. For students, the results from a summative assessment may determine what pathways are available to them in the future, and for teachers, the results are often subject to scrutiny by examination boards or used in benchmarking studies. In the future, the ability of universities to justify the quality of their assessment tasks in potential legal actions taken by disaffected students may also be necessary. Terms such as validity and reliability are frequently used in relation to assessment quality, yet many academics would be unsure how to measure these characteristics for their own assessment tasks despite the significant literature base on how to apply psychometric principles and statistical tools (Mislevy, Wilson & Chudowsky, 2002).

The design principles for preparing quality assessment tasks in higher education have been well documented (Biggs, 2002; Bull & McKenna, 2003; Case & Swanson, 2001; Dunn, Morgan, Parry & O'Reilly, 2004; James, McInnis & Devlin, 2002; McAlpine, 2002a; PASS-IT). There is also an extensive body of work in the discipline of validity and reliability testing for assessments and there are numerous descriptions that are readily available for academics on how to apply both psychometric principles and statistical analyses based on probability theories in the form of Classical Test Theory and Item Response Theory, particularly the Rasch Model (Baker, 2001; Downing, 2003; McAlpine, 2002b; Wright, 1977). Thus, there is no shortage of literature examples for academics to follow on preparing and analysing selected response questions; academics and academic developers should be in a position to continuously improve the quality of assessment tasks and student learning outcomes. However, the literature evidence for academics and academic developers generally using these readily available tools and theories is sparse (Knight, 2006).

Academics are generally not specialists in the research discipline of assessment, and they do not routinely analyze their assessments using the accepted standards associated with validity and reliability. Academics tend to rely on the accumulated discipline-based history about what constitutes an acceptable assessment standard, rather than attempt to apply quantitative principles from another discipline, especially if there is uncertainty about how to apply these principles appropriately. The key validation tool for the majority of assessments tends to rely on academic acumen rather than quantitative evidence (Knight, 2006; Price, 2005).

An analysis of the assessment in terms of acceptable standards of validity and reliability, as well as improvements that could be applied to the individual question items, their relative importance in an assessment and their ability to align with the stated objectives for the course, could lead to improvements in student performances. This analysis is particularly important for selected response items since they tend to be reused, either directly in subsequent years, or by rotating questions every few years. This is a common occurrence in academic practice, but is based more often on the time convenience afforded by the reuse of previously prepared questions, rather than a scholarly judgement about the efficacy of the question in discriminating between those students who have mastered concepts well and those who have not. Academic development units could assist in this endeavour by undertaking a routine analysis of assessment items using simple spreadsheet or database tools and preparing reports for academics that highlight the key issues that require attention, at least from the perspective of validity and reliability. This process could be used by academic developers to engage academics in the broader issue of using formative assessment to improve learning. Questions that have been shown to have good discrimination characteristics could be used during the learning stages, rather than just as discriminators in summative assessment tasks. Engaging academics with the analysis of assessment items is a potential pathway to engagement with methods to improve student learning, highlighting the importance of diagnostic and formative assessment coupled with appropriate feedback.

It could be argued that academics preparing high stakes summative assessments should develop the skills required to analyze the results of their assessments, but the reality often encountered in higher education institutions is that academics do not have the time, nor the incentives in place to allocate the time required to master these skills. By accepting that this situation is prevalent in most higher education institutions, the authors have proposed a relatively simple analysis and presentation format for reports on student assessment responses that academics could use to make judgments about their assessment tasks. We further posit that academic developers could then use these simple formats to engage academics in a discussion about the efficacy of using questions with particular characteristics in diagnostic and formative assessments in order to improve student learning outcomes.

Descriptions of basic and more sophisticated approaches to item analysis for academics have been reported but do not provide an adequate visual engagement component that would allow time-poor university staff to quickly determine the salient issues for a particular assessment (Kehoe, 1995; Maunder, 2002; Fowell, Southgate & Bligh, 1999).

Methodology

Survey on Analyzing Assessment Response and Results

We conducted an online survey of academics, predominantly from our Graduate Certificate in Higher Education program and the compulsory foundation program in university teaching. These groups were selected as they had recently been engaged in intense discussions about the relationship between assessment practices and student approaches to learning, and the importance of evaluation and reflection in the context of the scholarship of learning and teaching. The goals of the survey were to obtain feedback on participants' use of objective assessments, their awareness of psychometric principles and their knowledge of the statistical tools available for the analysis of student responses. Presentation formats illustrating different methods of reporting the analysis of multiple-choice question (MCQ) responses were discussed and the results from a number of MCQ tests for several disciplines were analyzed. Rasch analysis results, the use of the facility index, the discrimination index, the effectiveness of distracters and an item-person map analysis were all presented to the academic staff as valid analysis tools. The Rasch analysis and item-person map were generated using Winsteps®. All other data were presented using Excel's graphing tools.

Forty-five staff in total from across all major discipline types, including participants in the Graduate Certificate and foundation university teaching programs, and some staff known to be using MCQ assessments for large introductory classes, were contacted by email and invited to fill in the online survey.

Results and Discussion

We received twenty-one valid responses from the online survey (47% response rate), 77% indicated that they used MCQ assessments and 38% had undertaken some form of analysis of the student responses.

The results of the awareness and usage of psychometric principles and statistical tools for the analysis of MCQ assessments and their responses are summarised in Table 1. Academics were familiar with common statistical terms such as mean, median, standard deviation and percentiles. Some were familiar with the different types of terms used to describe validity, but very few were aware of the formal psychometric approaches associated with Classical Test Theory, the Rasch Model or Item Response Theory (Table 1).

Table 1: Summary of Academic Staff Responses to Awareness and use of Psychometric Principles and Statistical Tools for the Analysis of MCQ Response.

Which of the following assessment analysis terms are you familiar with? **n=21**		If you analyzed the student responses which of the following did you use? **n=8**	
Choice	**%**	**Choice**	**%**
Mean	100	Mean	38
Median	95	Median	33
Cronbach alpha coefficient	29	Cronbach alpha coefficient	10
Discrimination Index	19	Discrimination Index	10
Facility Index	5	Facility Index	5
Classical Test Theory	14	Classical Test Theory	10
Rasch Model	24	Rasch Model	5
Item Response Theory	24	Item Response Theory	5
Percentiles	90	Percentiles	19
Standard deviation	95	Standard deviation	33
Standard error of measurement	67	Standard error of measurement	10
Key	19	Weighted scoring	5
Options	24	Distribution of scores	15
Distracters	38	Item - Person Map	5
Construct validity	38	Distracter analysis	10
Angoff Method	10		
Content validity	48		

Only half of the respondents using MCQ assessments undertook some form of analysis of the student responses beyond simply reporting the student scores. Only 14% of the respondents who analyzed the student responses indicated that it influenced their teaching of the course, although 81% of all respondents indicated that an analysis of the student responses would be useful. Of the staff who analyzed their MCQ student responses, the mean (100%), median (88%), standard deviation (88%) and percentile (50%) were the most common properties used, reflecting a similar pattern to that observed for the whole group for the recognition of psychometric principles and statistical tools. Only one or two individuals used Classical Test Theory, the Rasch Model or Item Response Theory for analysis.

As part of the online survey, we presented academics with four different output file styles from both WinSteps® (Figure 1) and Excel. Table 2 summarizes the staff responses. The majority of staff found the presentation format for the output from a standard Winsteps® report not to be useful, although they were aware that it contained useful information that could be used to improve staff approaches to teaching. Typical open-ended comments from the survey responses for the standard Winsteps® format were:

> *Too complex for my needs and understanding*
>
> *Please tell me how many lecturers within the University would be able to meaningfully interpret the data above in relation to their students' learning outcomes?*
>
> *It would help me analyse the questions and give useful information (maybe some of these features in a graphical format would be more used though)*

TABLE 10.1 ZOU721ws.txt
INPUT: 146 persons, 30 items MEASURED: 146 persons, 30 items, 2 CATS 3.57.2

person: REAL SEP.: 1.57 REL.: .71 ... item: REAL SEP.: 5.00 REL.: .96

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	ESTIM DISCR	P-VALUE	item
26	25	145	70.35	2.35	1.19	1.2	1.41	1.9	A .03	.79	.17	I0026
23	82	146	49.53	1.77	1.11	2.0	1.23	2.4	B .07	.41	.56	I0023
29	64	146	55.14	1.78	1.14	2.3	1.21	2.5	C .05	.38	.44	I0029
21	90	146	46.99	1.80	1.04	.8	1.18	1.6	D .14	.73	.62	I0021
30	73	145	52.18	1.77	1.12	2.2	1.15	1.7	E .07	.40	.50	I0030
4	55	146	58.06	1.83	1.07	1.0	1.10	1.1	F .15	.78	.38	I0004
17	36	146	65.11	2.06	1.01	.2	1.09	.6	G .22	.96	.25	I0017
18	74	146	52.01	1.76	1.03	.6	1.05	.6	H .19	.84	.51	I0018
1	76	146	51.39	1.76	1.01	.3	1.03	.4	I .21	.92	.52	I0001
7	102	144	42.55	1.92	1.01	.2	.95	-.3	J .18	1.00	.71	I0007
19	93	145	45.78	1.82	1.01	.2	.97	-.2	K .20	.99	.64	I0019
27	19	146	74.07	2.62	1.01	.1	.94	-.2	L .23	1.00	.13	I0027
20	106	146	41.48	1.94	1.01	.1	.94	-.3	M .19	1.01	.73	I0020
22	79	146	50.46	1.76	1.00	.0	.99	-.1	N .23	1.00	.54	I0022
28	63	146	55.46	1.79	1.00	.0	.97	-.3	O .25	1.04	.43	I0028
6	108	146	40.71	1.97	.99	-.1	.97	-.1	o .19	1.02	.74	I0006
8	114	145	38.03	2.10	.98	-.1	.93	-.3	n .19	1.04	.79	I0008
9	121	143	33.60	2.38	.97	-.1	.86	-.5	m .20	1.04	.85	I0009
2	104	146	42.22	1.91	.97	-.4	.92	-.5	l .24	1.09	.71	I0002
3	134	146	26.23	3.06	.97	.0	.74	-.6	k .19	1.04	.92	I0003
13	57	145	57.27	1.82	.97	-.5	.96	-.5	j .29	1.10	.39	I0013
10	46	141	60.91	1.92	.95	-.5	.97	-.3	i .29	1.09	.33	I0010
11	73	145	52.21	1.77	.97	-.7	.94	-.7	h .28	1.19	.50	I0011
15	110	145	39.64	2.02	.96	-.4	.89	-.6	g .24	1.08	.76	I0015
12	98	146	44.34	1.85	.96	-.6	.92	-.6	f .26	1.13	.67	I0012
16	106	146	41.48	1.94	.96	-.5	.91	-.5	e .25	1.10	.73	I0016
5	83	145	49.08	1.78	.94	-1.0	.91	-.9	d .30	1.28	.57	I0005
14	78	146	50.77	1.76	.94	-1.2	.89	-1.2	c .32	1.34	.53	I0014
25	50	145	59.71	1.87	.88	-1.5	.85	-1.6	b .42	1.30	.34	I0025
24	70	146	53.25	1.77	.85	-2.9	.80	-2.5	a .45	1.72	.48	I0024
MEAN	79.6	145.4	50.00	1.96	1.00	.0	.99	.0				
S.D.	27.8	1.1	10.21	.29	.07	1.1	.14	1.1				

Figure 1: WinSteps® Output for a MCQ Summative Test

Table 2: Summary of academic staff responses to awareness and use of psychometric principles and statistical tools for the analysis of MCQ responses.

Winsteps® format useful, Figure 1, n = 21		Discrimination Index format useful, Figure 4, n = 21		Distracter format useful, Figure 5, n = 21		Person-item map format useful Figure 6, n = 20	
Yes	No	Yes	No	Yes	No	Yes	No
38%	62%	76%	24%	57%	43%	70%	30%

We have proposed a number of simple presentation formats for MCQ analysis reports that we believe highlight the key features that academics should engage with, and which will have a positive impact on staff teaching and student learning. From our discussions with academics in our graduate certificate and foundation teaching programs, we were aware that academic development units will have more impact if they provide information to staff in a simple, easily understood format, rather than solely concentrating on development workshops or seminars of a general nature (Prebble, Hargraves, Leach, Naidoo, Suddaby & Zepke, 2005; Prosser, Rickinson, Bence, Hanbury & Kulej, nd).

Score Distribution

The score distribution and overall mean score from our WinSteps® example is shown in a column graph format in Figure 2. McAlpine (2002a) and Johnstone (2003) have suggested that an acceptable mean mark across assessments conducted in norm-referenced modes should be between 50-60%, indicating that on this basis at least this MCQ assessment was acceptable. We can see immediately that the distribution is centred approximately as expected for this norm-referenced activity, and that the minimum scores are above that apparently expected from purely statistical guessing. Burton & Miller (1999) and Burton (2001) have described methods for quantifying the effects of chance on the '50/60% overlap' score region, and have shown that the impact of guessing correct responses in 4 and 5 option MCQ tests can reduce test reliability significantly. Without the use of negative or confidence level marking (McCabe & Barrett, 2003), it is difficult to separate purely statistical guessing from 'informed guessing'.

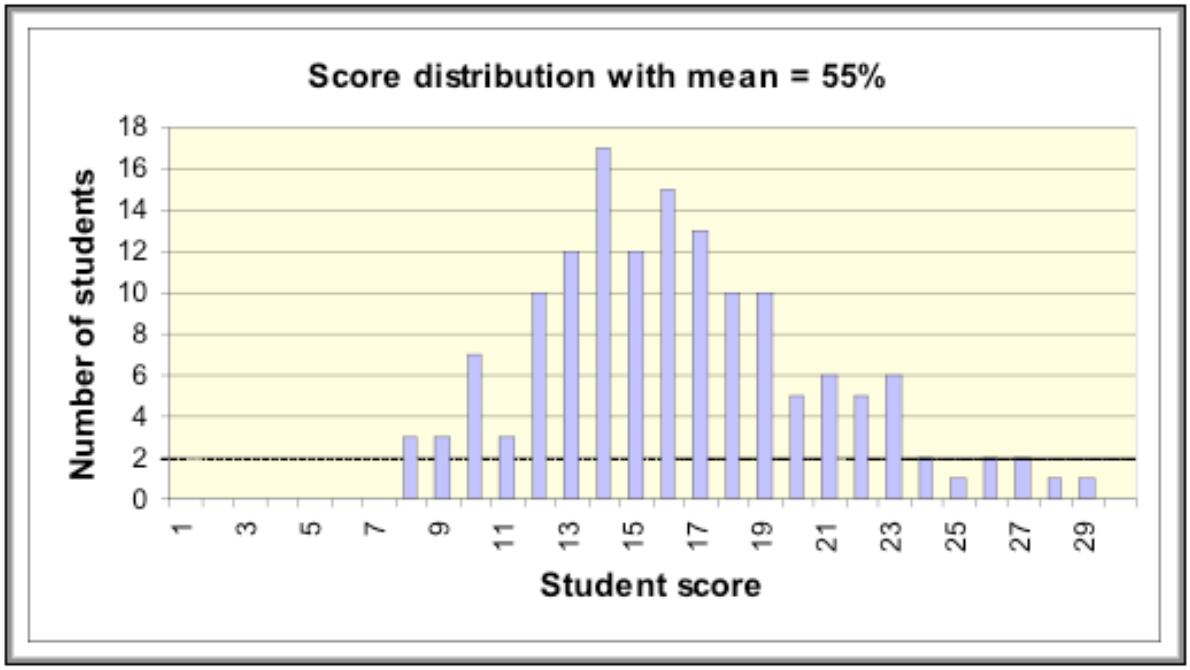

Figure 2: Score Distribution for a MCQ Summative Test, Consisting of 30 Items with 5 Options Each, Undertaken by 146 students.

Facility Index

Classical Test Theory (McAlpine, 2002b) can be used to determine a facility index (FI) for each item in the test. The FI indicates how many students chose the designated correct response compared to those who chose other options (or distracters), and is expressed as a fraction. Johnstone (2003) and McAlpine (2002b), have suggested that academic staff should aim for a FI of between 0.3 – 0.8 for each question. Figure 3 illustrates a column graph showing the FI for each item in our 30-question MCQ example. Academic staff should be made aware of any question which returns a FI outside the suggested 0.3 – 0.8 range (as indicated by the darker shading in Figure 3). FI values for items that are above 0.8 indicate that most students selected the designated correct response to the item, whilst FI values below 0.2 indicate that only a few students chose the correct response. We can see that questions 3 and 9 had a FI above 0.8 and questions 17, 26 and 27 had a FI below 0.3. This does not automatically mean that these questions should be removed from the assessment, or that they were inappropriate questions. For academic staff the priority would be to review these questions in particular and decide if they were consistent with the stated objectives for the assessment and allowed students to demonstrate learning and skill development. Difficult questions may have been intentionally incorporated into the test.

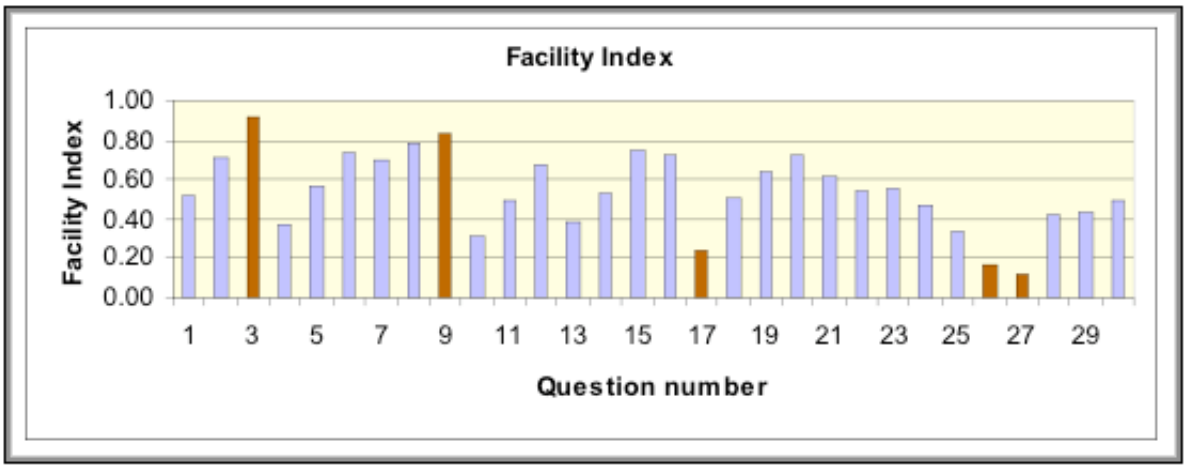

Figure 3: Facility Index Graph for the Example MCQ Summative Test.

The order in which questions are presented in a test has been shown to have minimal influence on the overall scores obtained by students (McLeod, Zhang & Hao, 2003). Questions may be arranged according to topic, similarity of concepts, difficulty order or simply at random. In order to assist students in gaining confidence in answering questions it is often beneficial to commence with relatively 'easy' questions, and gradually increase the difficulty level as the test proceeds (Clariana & Wallace, 2002; Haladyna, Downing & Rodriguez, 2002). We can see from Figure 3 that the FI values tended to decrease for the second half of the 30 questions, although the trend is not uniform. This is useful information for academic staff who may wish to arrange questions in a particular order.

Discrimination Index

The discrimination index (DI) is another very simple indicator that may be used to measure the ability of a question to differentiate between high and low achieving students (McAlpine, 2002b). The DI for each question can be calculated by subtracting the FI for each question for the bottom third of the class from the FI for each question for the top third of the class (ranked according to their overall score on the assessment). There are more sophisticated methods for calculating the DI but the method described here is simple and useful for most academic staff in universities (McAlpine, 2002b). The DI may range from 1.0 to –1.0, with 1.0 being a perfect correlation between students selecting the correct response and also scoring high marks on the test and –1.0 being for questions where students answered incorrectly but scored highly overall. Typical values recommended for the DI value are above 0.3 (Johnstone, 2003 and McAlpine, 2002b). In Figure 4 we can see that all questions in the test had a positive DI, meaning that students who answered each question correctly, also scored

more highly overall. However, many of the questions had DI values below the suggested 0.3, indicating that they did not allow much discrimination between students, since high achieving and low achieving students answered the questions equally well. In particular, we can see that questions 26, 27, 29 and 30 had low DI values but appeared at the end of the test, a position where more difficult and discriminating questions might have been expected. Questions 26 and 27 have both a low discrimination and facility index, suggesting that these questions may not be appropriate.

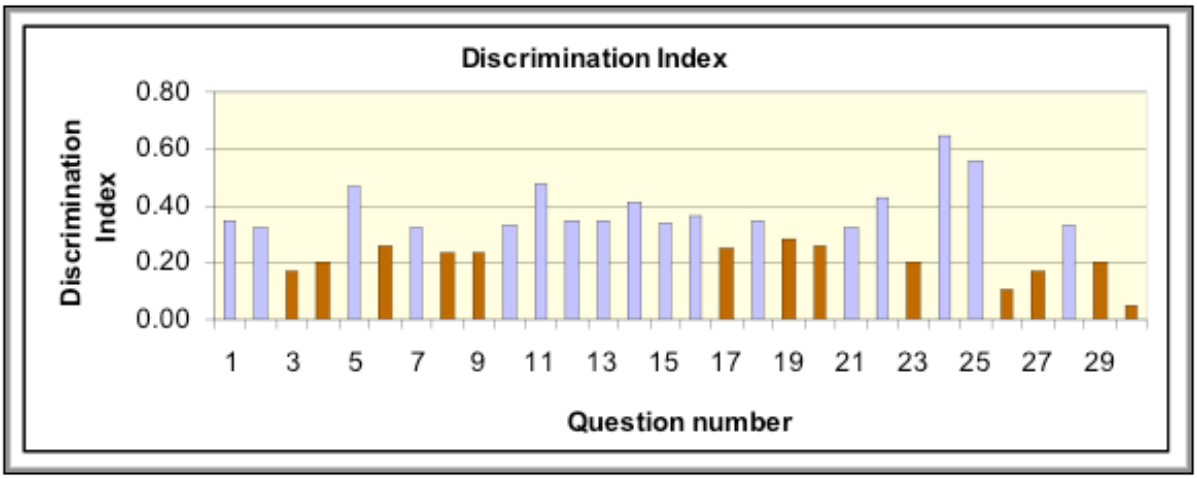

Figure 4: Discrimination Index Plot for the Example MCQ Summative Test

Figure 4 was presented as part of the staff online survey, and Table 2 indicates that a significant majority of the participants found this presentation format useful. Open-ended comments from staff included the following, with the first comment likely indicating a misunderstanding of the use of the DI:

> *This would help me see which questions were difficult and easy and match that to the learning objective.*
>
> *A question that received low value would indicate a topic that was confusing to a large portion of the class (both high & low performers) - could indicate an error in the material presented or that the material wasn't well presented/clear.*
>
> *I don't see how a correlation between individual questions and overall performance could improve the MCQs.*

Distracters

Distracters (incorrect responses) are designed to differentiate students who have learnt the material from those who have not. Useful distracters would normally cover a known misconception experienced by previous students, and factual errors that are familiar to the teacher, and should have a student response value of at least 20-30% for each distracter. It is wasteful of the academics' and students' time to add distracters to an item that have very low student response values.

The example data provided here were based on a series of 5 option MCQs consisting of one key and 4 distracters. Figure 5 illustrates how students responded to each of the options in the assessment, shaded differently for the responses A, B, C, D and E. This representation provides teaching staff with a significant amount of useful data. A quick visual inspection informs us that option E is underrepresented. This would imply that the assessor has not used E as a key for many questions, or that thinking up 5 options was difficult and E was usually assigned to an option that was not as plausible as the other options. For each question, the percentage of responses for each option can be visually determined by examining the extent of the shaded bar. If assessors have decided to commence the assessment with easier items, and gradually increase the level of difficulty of items as the test proceeds, then items in the second half of the assessment would often be expected to have a more even distribution of student responses to each option, assuming that later questions have a higher DI for each option. The shading pattern in Figure 5 would be expected to change as we proceeded from left to right. This type of quick visual overview enables academics to focus on key issues for their assessment structure and the resulting outcomes.

Thus for question 3 we can see that students did not choose the distracters over the key often. This is the type of question that the academic should review for the efficacy of the options. Questions 25-30 all have good selections across all options. This may mean that the options were testing known misconceptions, or that students had little idea what the correct response should be; the DI correlation could then be used to determine whether high achieving students were choosing the key.

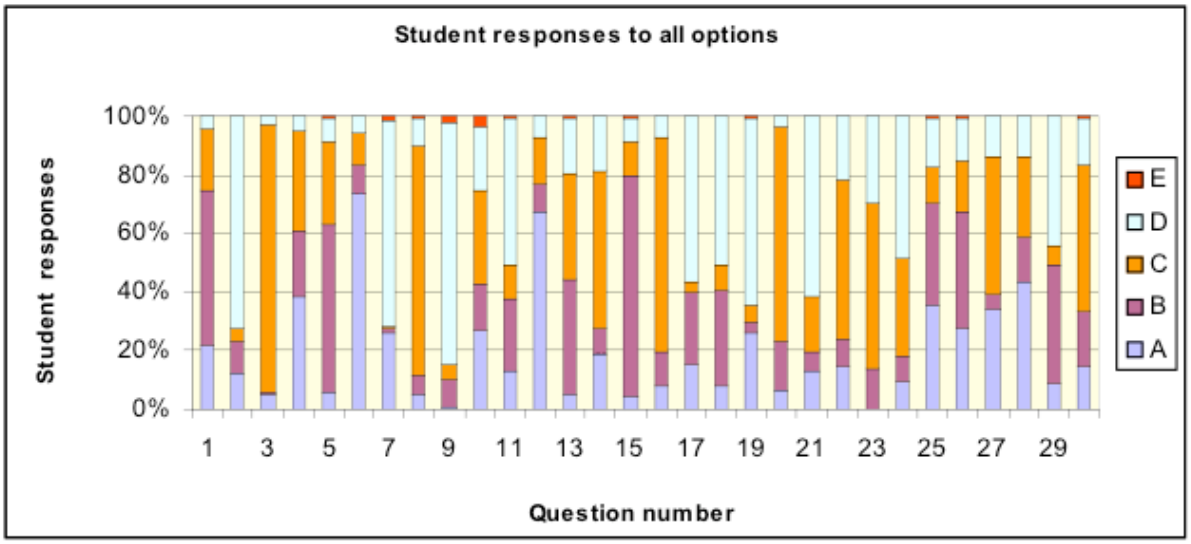

Figure 5: Student Responses to all Options for the Example MCQ Summative Test.

Figure 5 was presented to academic staff as part of the online survey, and received mixed responses as to its usefulness with an equal split between those academics who found it helpful, and those who were not clear what it represented (Table 2). Some of the open-ended comments from survey participants included:

> *Too complex for my needs and understanding.*
>
> *Provides at a glance, questions where majority of students are responding to determined best answer and other questions where the distribution might be more even across options, suggesting the question has problems with it or if distribution shows many responding to incorrect answer, then students need assistance or there is a mistake in the answer.*
>
> *I am afraid that my eyes glazed over when I looked at the chart above. I have taught students in social research and program evaluation courses (in three Universities over ten years) to only include charts in their outputs that present the data clearly.*
>
> *A chart is less useful for this than seeing a list of the answers with the percentage of students selecting those answers. I don't think having it in a bar format adds anything and actually I find it more confusing than just answers with percentages next to them. For example in the chart above in question one B ranges from approx 22% to 77%. This, I assume, means that approx 55% of students chose B. But it's easier to just see B 55% to get this information than the bars which can be harder to distinguish differences of say 10% visually (which I would consider significant).*

Academics often have different preferences for the way material is presented, just as students have different preferred learning styles. The data presented in bar graph format in Figure 5 could also be presented as a table, with the percentage responses to each option and an indication of the key for each item (Table 3). The information contained in Figure 5 or Table 3 could be used by either academics or academic developers to not only improve the items in this particular assessment, but also to develop diagnostic and formative assessment tasks that could potentially improve student learning. Items that display good DI values could be used early in the teaching period as formative tasks that direct student learning by providing appropriate feedback. This process of identifying assessment items with good discrimination characteristics could be used proactively by academics to improve student learning on an iterative basis.

Table 3: Summary of student responses to MCQ assessment as outlined in Figure 5.

%	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Correct response	B	D	C	A	B	A	D	C	D	C	D	A	B	C	B
A	22	12	5	38	6	74	26	5	1	27	13	67	5	18	4
B	52	11	1	23	57	10	1	7	10	16	24	10	39	9	76
C	21	4	92	34	28	10	1	78	5	32	12	16	36	54	12
D	4	71	3	5	8	6	70	10	83	22	50	7	19	18	8
E	0	0	0	0	1	0	1	1	3	3	1	0	1	0	1
FI	0.52	0.71	0.92	0.37	0.57	0.74	0.70	0.78	0.83	0.31	0.50	0.67	0.39	0.53	0.75
DI	0.35	0.32	0.17	0.20	0.47	0.26	0.32	0.24	0.24	0.33	0.47	0.34	0.35	0.41	0.34

	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30
Correct response	C	B	D	D	C	D	C	C	D	B	C	D	A	D	C
A	8	15	8	26	7	13	14	0	10	35	27	34	44	9	8
B	12	25	33	4	16	7	10	14	9	35	40	5	15	40	12
C	73	3	8	5	73	18	54	56	33	12	18	48	28	7	73
D	7	56	51	64	3	62	22	30	48	17	14	14	14	44	7
E	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0
FI	0.72	0.24	0.50	0.63	0.72	0.61	0.54	0.56	0.48	0.34	0.17	0.12	0.43	0.43	0.50
DI	0.36	0.25	0.35	0.28	0.26	0.32	0.43	0.20	0.64	0.56	0.10	0.17	0.33	0.20	0.05

FI represents Facility Index data from Figure 3; DI represents Discrimination Index data from Figure 4.

Reliability and Validity

How does an academic know if the overall assessment is valid and reliable? A valid assessment is one that measures what it professes to measure; a reliable assessment is one that would produce similar results over a period of time when used by students of similar ability and in the same circumstances. The most common measure used for internal consistency within a single test is the correlation between items using a Cronbach's Alpha (α) coefficient of reliability, and although there have been discussions in the literature over whether this single measure is a true indicator of reliability for MCQ assessments (Burton, 2004), when used in conjunction with the other data suggested in this paper, it will give academic staff a reasonable summary of a particular assessment. McAlpine (2002b) has suggested MCQ assessments should aim for an α of approximately 0.70. The WinSteps® data from our example indicate an α of 0.69 and this can simply be presented as a numerical value to academic staff.

Overall summary of MCQ assessment

Figure 6 illustrates the person-item map from a WinStep® output for our MCQ assessment example. This shows the distribution of scores as a percentage on the left (similar information to that presented in graphical format in Figure 2), with groups of 2 student scores indicated by a # and 1 student score represented by a period(.). The question (item) number appears on the right hand side (where I0003 represents item 3), arranged with the easiest (highest FI values) on the bottom and the most difficult (in this example item 27 represented by I0027) on the top. The letter M indicates the position of the mean, S one standard deviation from the mean and T indicating two standard deviations from the mean. It can be seen that questions 3, 8 and 9 were answered correctly by most students since they are more than one standard deviation away from the mean (M and S on the right hand side), and academics could decide whether these question are relevant in a norm-referenced assessment. Questions 10, 17, 26 and 27 were quite difficult, and again academics could decide if this was appropriate.

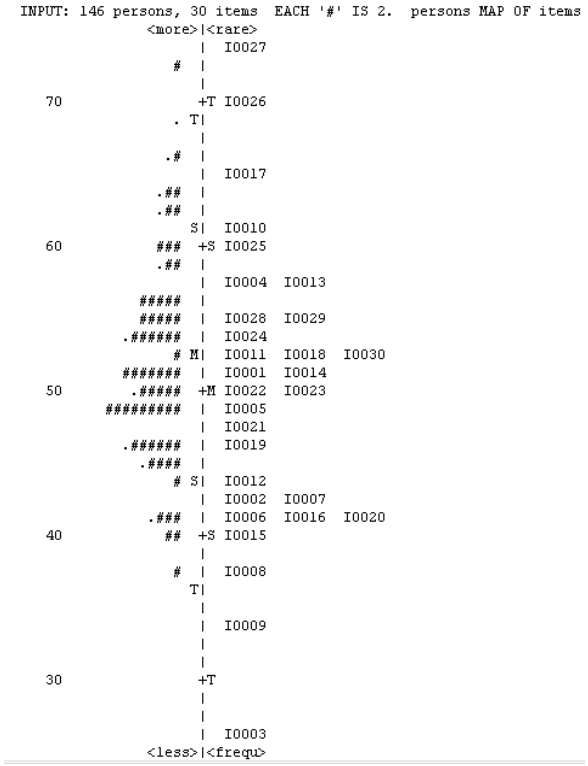

Figure 5: Student Responses to all Options for the Example MCQ Summative Test in the Form of a Person-Item Map.

The purpose of this diagram is to provide academics with a visual summary of both the performance of the class as a whole (from the distribution of scores on the left hand side) and the level of difficulty of each question, all on one diagram. The distribution of student scores and difficulty level of items along the vertical axis provides a visual check on unexpected distribution patterns. The same information can be presented as individual tables of figures, or as separate bar graphs, but the advantage of the person-item map is that is presents in a relatively simple manner a large amount of analytical data. Such data presentation could be undertaken by academic development units as a service to areas using MCQ assessment items. By providing this service, academic developers would be in a position to engage academics in the broader discussion about their assessment strategies, especially the use of diagnostic and formative assessment for improving student learning. These discussions would be framed with an evidenced-based approach and academics would be able to track the efficacy of their activities.

Figure 6 was presented in the online survey and overall was regarded as useful by the majority of participants, although not all thought it would assist them (Table 2). Open-ended comments from academic staff about this presentation format included:

> *Helpful in visually seeing questions that really made students think. Perhaps this could be designed to be more visual - colour?*
>
> *This type of representation shows how well the distribution of student scores matches the difficulty of the question and would be more the type of information I am after.*
>
> *Again this is a standard output stem and leaf plot which takes some work - there are much better ways to present these kinds of data.*

Summary

It is common practice for academic staff to set an assessment, mark it, report the students' scores and then give the assessment no further thought until the next iteration of the cycle. This is understandable when academics are pressured to report students' scores as soon as possible so that grades and graduations can be finalized and other activities such as research and new course designs continually demand attention. Academics participating in our graduate certificate and foundation teaching program have indicated that academic development units could assist them by preparing appropriate reports on the analysis of the student responses and scores from their assessments in a succinct and visually engaging manner. They are prepared to allocate time to reflect on student responses and how they could improve their teaching and assessment designs. What they require are reports that can be quickly interpreted and where the issues that will have the greatest impact on student performance or outcomes are highlighted. Academic developers could further use the process of presenting the results of a MCQ assessment analysis to engage academics in the use of specific items for diagnostic and formative assessment, thus providing an evidence-based pathway for the improvement of student learning.

We have presented a series of simple visual representations of psychometric or statistical data derived from the analysis of MCQ assessments as a first stage in this reporting process. We are not proposing that the particular formats presented in this paper are necessarily ideal. There are many adaptations that may highlight the key features arising from student responses more effectively, but what this paper has demonstrated is that academic development units need to reflect on how they may assist academics to improve student performance and learning outcomes in a practical way. Presenting academics with simple (but not simplistic) analytical tools and easily understood frameworks will facilitate engagement with the underlying theoretical and pedagogical issues related to assessment and student learning.

References

Baker, F. B. (2001) *The Basics Of Item Response Theory*. Retrieved from: http://edres.org/irt/baker (accessed 6 December 2007).

Biggs, J.B. (2002) *Aligning teaching and assessment to curriculum objectives*. LTSN Imaginative Curriculum Guide IC022 Retrieved from: http://www.heacademy.ac.uk/ourwork/learning/assessment (accessed 6 December 2007).

Bull, J. & McKenna, C. (2003) *Blueprint for Computer-assisted Assessment*. (London, Routledge Falmer).

Burton, R. F. & Miller, D. J. (1999) Statistical Modelling of Multiple-choice and True/False Tests: ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, 24(4), 399-411.

Burton, R. F. (2001) Quantifying the Effects of Chance in Multiple Choice and True/False Tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26(1), 41-50.

Burton, R. F. (2004) Multiple choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education*, 29(5), 585-595.

Case, S.M. & Swanson, D.B. (2001) *Constructing written test questions for the basic and clinical sciences* (3rd Edn). National Board of Medical Examiners USA.

Clariana, R & Wallace, P. (2002) Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593–602.

Downing, S. M. (2003) Item response theory: applications of modern test theory in medical education. *Medical Education*, 37, 739–745.

Dunn, L., Morgan, C., Parry, S. & O'Reilly, M. (2004) *The Student Assessment Handbook: New Directions in Traditional and Online Assessment*. (London, Routledge Falmer).

Fowell, S. L., Southgate, L. J. & Bligh, J. G. (1999) Evaluating assessment: the missing link? *Medical Education*, 33, 276-281.

Haladyna, T. M., Downin, S. M. & Rodriguez, M. C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education,* 15(3), 309-333.

Jackson, T. R., Draugalis, J. R., Slack, M. K., Zachry, W. M. & D'Agostino, J. (2002) Validation of Authentic Performance Assessment: A Process Suited for Rasch Modeling. *American Journal of Pharmaceutical Education*, 66, 233-243.

James, R., McInnis, C. & Devlin, M. (2002) *Assessing Learning in Australian Universities: Ideas, strategies and resources for quality in student assessment*. Centre for the Study of Higher Education, The University of Melbourne and The Australian Universities Teaching Committee, Canberra, Australia. Retrieved from: http://www.cshe.unimelb.edu.au/assessinglearning (accessed 6 December 2007).

Johnstone, A. (2003) *LTSN Physical Sciences Practice Guide Effective Practice in Objective Assessment The Skills of Fixed Response Testing*. Retrieved from: http://www.heacademy.ac.uk/physsci/home/pedagogicthemes/assessment (accessed 6 December 2007).

Kehoe, J. (1995) Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10).

Knight, P. (2006) The local practices of assessment. *Assessment & Evaluation in Higher Education*, 31(4), 435-452.

McAlpine, M. (2002a) *Principles of Assessment* (ed CAA Centre, University of Luton), Retrieved from: http://caacentre.lboro.ac.uk/dldocs/Bluepaper1.pdf (accessed 6 December 2007).

McAlpine, M. (2002b) *A Summary of Methods of Item Analysis* (ed CAA Centre, University of Luton), Retrieved from: http://caacentre.lboro.ac.uk/dldocs/Bp2final.pdf (accessed 6 December 2007).

McCabe, M. and Barrett, D. (2003), 'It's a MUGS game! Does the mathematics of CAA matter in the computer age?' Maths CAA Series. Retrieved from: http://ltsn.mathstore.ac.uk/articles/maths-caa-series/apr2003/index.shtml (accessed 6 December 2007).

McLeod, I., Zhang, Y. & Hao Yu, H. (2003) Multiple-Choice Randomization, *Journal of Statistics Education*, 11(1), Retrieved from: http://www.amstat.org/publications/jse/v11n1/mcleod.html (accessed 6 December 2007).

Maunder, P. (2002) In support of multiple choice questions: some evidence from Curriculum 2000, Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002. Retrieved from: http://www.leeds.ac.uk/educol/documents/00002575.htm (accessed 6 December 2007).

Mislevy, M. R., Wilson, K. E. & Chudowsky, N. (2002) *Psychometric Principles in Student Assessment*. In T. Kellaghan & D. Stufflebeam (Eds.), International Handbook of Educational Evaluation (pp. 489-531). (Netherlands, Kluwer Academic Press).

Price, M. (2005) Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education*, 30(3), 215-230.

Prosser, M, Rickinson, M., Bence, V., Hanbury, A. & Kulej, M. nd. Formative evaluation of accredited programmes. The Higher Education Academy Retrieved from: http://www.heacademy.ac.uk/researchpublications.htm (accessed 6 December 2007)

PASS-IT *Good Practice Guide in Question and Test Design, Project on Assessment in Scotland – using Information Technology*. Retrieved from: http://www.pass-it.org.uk/resources/031112-goodpracticeguide-hw.pdf (accessed 6 December 2007).

Prebble, T., Hargraves,H., Leach, L., Naidoo, K., Suddaby, G. & Zepke, N. (2005) Academic Staff Development: A summary of a synthesis of research on the impact of academic staff development programmes on student outcomes in undergraduate tertiary study. Rivers, J (Ed) Retrieved from: http://www.educationcounts.govt.nz/__data/assets/pdf_file/0020/7319/academic-staff-development-summary.pdf (accessed 6 December 2007).

Winstep. Retrieved from: http://www.winsteps.com (accessed 6 December 2007).

Wright, B. D. (1977) Solving Measurement Problems with the Rasch Model, *Journal of Educational Measurement*, 14, 97-116.

Please cite as:

Crisp, G. & Palmer, E. (2007). Engaging Academics with a Simplified Analysis of their Multiple-Choice Question (MCQ) Assessment Results. *Journal of University Teaching and Learning Practice*, 4(2), 88-106. http://jutlp.uow.edu.au/2007_v04_i02/pdf/crisp.pdf